



IPN

Atty. Dkt. No. 089367-0127

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Fumitoshi MIZUTANI, et al.
Title: DATA PROCESSING APPARATUS AND DATA PROCESSING METHOD
Appl. No.: 10/827,433
Filing Date: 04/20/2004
Examiner: Unknown
Art Unit: Unknown

CLAIM FOR CONVENTION PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

The benefit of the filing date of the following prior foreign application filed in the following foreign country is hereby requested, and the right of priority provided in 35 U.S.C. § 119 is hereby claimed.

In support of this claim, filed herewith is a certified copy of said original foreign application:

Japanese Patent Application No. 2003-115621
filed 04/21/2003.

Respectfully submitted,

RN 38,072

Date: May 26, 2004

FOLEY & LARDNER LLP
Customer Number: 22428
Telephone: (202) 672-5407
Facsimile: (202) 672-5399

By David A. Blumenthal

for David A. Blumenthal
Attorney for Applicant
Registration No. 26,257

US

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 4 月 2 1 日
Date of Application:

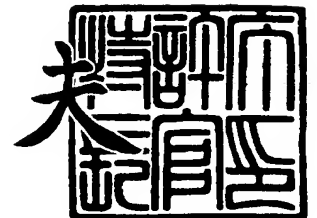
出 願 番 号 特 願 2 0 0 3 - 1 1 5 6 2 1
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 1 1 5 6 2 1]

出 願 人 日 本 電 気 株 式 有 限 公 司
Applicant(s):

2 0 0 4 年 3 月 2 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 1 5 7 0 2

【書類名】 特許願

【整理番号】 62703096

【提出日】 平成15年 4月21日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 15/16
G06F 11/16

【発明者】

【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

【氏名】 水谷 文俊

【発明者】

【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

【氏名】 尾田 眞也

【特許出願人】

【識別番号】 000004237

【氏名又は名称】 日本電気株式会社

【代理人】

【識別番号】 100095407

【弁理士】

【氏名又は名称】 木村 満

【手数料の表示】

【予納台帳番号】 038380

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9715824

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 データ処理装置

【特許請求の範囲】

【請求項 1】

同一のデータ送信元から同一のデータを受信する複数の受信インタフェース部を備え、前記複数の受信インタフェース部が受信したデータの処理を並列して行うデータ処理装置において、

前記各受信インタフェース部は、受信したデータにエラーが発生すると、データの受信を停止し、異なる受信インタフェース部に、前記データ送信元からのデータ受信を停止させる通信エラー信号を出力して、前記データ送信元にデータの再送を要求する通信エラー処理部を備えた、

ことを特徴とするデータ処理装置。

【請求項 2】

前記各受信インタフェース部の通信エラー処理部は、受信したデータの一部にエラーが発生すると、エラーが発生したデータを破棄して、前記データ送信元に、破棄したデータの再送を要求するように構成されたものである、

ことを特徴とする請求項 1 に記載のデータ処理装置。

【請求項 3】

前記データ送信元は、同一のシリアルデータを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したシリアルデータにエラーが発生すると、エラーが発生したシリアルデータ及び当該データに続いて受信したシリアルデータを破棄して、前記データ送信元に破棄したシリアルデータの再送を要求するように構成された、

ことを特徴とする請求項 1 に記載のデータ処理装置。

【請求項 4】

前記データ送信元は、各パケットにシーケンス番号を付加してパケット単位で前記データを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したパケットのデータにエラーが発生すると、前記受信した各パケットに付加されたシーケンス番号

に基づいてパケット単位で、前記データ送信元に、データの再送を要求するように構成されたものである、

ことを特徴とする請求項 1 乃至 3 のいずれか 1 項に記載のデータ処理装置。

【請求項 5】

所定のクロック信号の周波数を分周して同期信号を生成し、生成した同期信号を前記各受信インタフェース部に供給する分周器を備え、

前記各受信インタフェース部は、前記分周器が供給した同期信号に従ってデータを受信するものである、

ことを特徴とする請求項 1 乃至 4 のいずれか 1 項に記載のデータ処理装置。

【請求項 6】

送信対象のデータを同じタイミングで複数のデータ送信先に送信する送信インタフェース部を備えたデータ処理装置において、

前記送信インタフェース部は、送信対象のデータを、所定のクロック信号の 1 周期以内で送信可能なデータ長のデータに分割してパケットデータを生成し、生成した各パケットデータを前記クロック信号に同期させて同じタイミングで前記複数の送信先に送信するように構成された、

ことを特徴とするデータ処理装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、同一のデータを並列して処理するデータ処理装置に関する。

【0002】

【従来の技術】

データ処理を実行するコンピュータシステムとして、既存のコンポーネントを流用して、冗長化された構成のフォールトトレラントコンピュータシステムがある（例えば、特許文献 1 参照）。このコンピュータシステムでは、ロックステップ方式が採用されている。

【0003】

このロックステップ方式とは、冗長化構成の複数のプロセッサを備え、複数の

プロセッサが同一のデータを同期して並列処理し、複数のプロセッサからの出力を比較してエラーを検出すると、そのエラーを修復するようにした方式である。

【0004】

また、近年のコンピュータシステムにおいては、プロセッサと I/O システムとの間の接続には、PCI-Express、Hyper-Transport（登録商標）、InfiniBand（登録商標）等のような高速でデータの送受信を行える高速シリアルリンク方式が採用されつつある。

【0005】

【特許文献 1】

特開平 9-128349 号公報（第 5-7 頁、図 1）

【0006】

【発明が解決しようとする課題】

しかし、従来の冗長構成のコンピュータシステムにおいて、このような高速のデータ送受信方式を採用することにより、データの送受信速度が速くなると、複数のプロセッサが処理するデータの同一性を保証することが難しくなり、通信エラーも生じやすくなる。

【0007】

プロセッサと I/O システムとの間でデータを送受信するインタフェース部は、通信エラーを検出すると、それぞれ、異なるタイミングでデータの再送信を要求してしまう。各インタフェース部が、データの再送信を要求すると、各プロセッサでの処理のタイミングや順序にずれが生じ、ロックステップ方式を維持できなくなり、複数のプロセッサが同一のデータ処理を同期して行うことが難しくなる。

【0008】

また、このようなコンピュータシステムにおいては、通信線の配線長によってデータ遅延が生じやすい。データ遅延が生じて複数のプロセッサの処理タイミングがずれると、同じように複数のプロセッサが同一のデータ処理を同期して行うことが難しくなる。このため、等長配線を厳密に行わなければならない、システムの筐体構造やボード設計、ボード構成の自由度にも大きな制約が生じることにな

る。

【0009】

本発明は、このような従来の問題点に鑑みてなされたもので、通信エラーが生じた場合でも同一のデータ処理を同期して行うことが可能なデータ処理装置を提供することを目的とする。

【0010】

【課題を解決するための手段】

この目的を達成するため、本発明の第1の観点に係るデータ処理装置は、同一のデータ送信元から同一のデータを受信する複数の受信インタフェース部を備え、前記複数の受信インタフェース部が受信したデータの処理を並列して行うデータ処理装置において、

前記各受信インタフェース部は、受信したデータにエラーが発生すると、データの受信を停止し、異なる受信インタフェース部に、前記データ送信元からのデータ受信を停止させる通信エラー信号を出力して、前記データ送信元にデータの再送を要求する通信エラー処理部を備えたものである。

【0011】

このような構成によれば、通信エラーが生じた場合でも同一のデータ処理を同期して行うことが可能になる。

【0012】

前記各受信インタフェース部の通信エラー処理部は、受信したデータの一部にエラーが発生すると、エラーが発生したデータを破棄して、前記データ送信元に、破棄したデータの再送を要求するように構成されたものであってもよい。

【0013】

前記データ送信元は、同一のシリアルデータを送信するものであって、前記各受信インタフェース部の通信エラー処理部は、受信したシリアルデータにエラーが発生すると、エラーが発生したシリアルデータ及び当該データに続いて受信したシリアルデータを破棄して、前記データ送信元に破棄したシリアルデータの再送を要求するように構成されたものであってもよい。

【0014】

前記データ送信元は、各パケットにシーケンス番号を付加してパケット単位で前記データを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したパケットのデータにエラーが発生すると、前記受信した各パケットに付加されたシーケンス番号に基づいてパケット単位で、前記データ送信元に、データの再送を要求するように構成されたものであってもよい。

【0 0 1 5】

所定のクロック信号の周波数を分周して同期信号を生成し、生成した同期信号を前記各受信インタフェース部に供給する分周器を備え、

前記各受信インタフェース部は、前記分周器が供給した同期信号に従ってデータを受信するものであってもよい。

【0 0 1 6】

本発明の第 2 の観点に係るデータ処理装置は、

送信対象のデータを同じタイミングで複数のデータ送信先に送信する送信インタフェース部を備えたデータ処理装置において、

前記送信インタフェース部は、送信対象のデータを、所定のクロック信号の 1 周期以内で送信可能なデータ長のデータに分割してパケットデータを生成し、生成した各パケットデータを前記クロック信号に同期させて同じタイミングで前記複数の送信先に送信するように構成されたものである。

【0 0 1 7】

【発明の実施の形態】

以下、本発明の実施の形態に係るデータ処理装置を図面を参照して説明する。

尚、本実施の形態に係るデータ処理装置を、冗長構成を有するコンピュータシステムとして説明する。

本実施の形態に係るコンピュータシステムの構成を図 1 に示す。

本実施の形態に係るコンピュータシステムは、冗長化構成の複数のプロセッサを備え、複数のプロセッサが同一のデータを同期して並列処理するロックステップ方式に従って動作するフォールトトレラントコンピュータシステムであり、サブシステム 1、2 を備えて構成される。

【0018】

サブシステム 1 は、演算システム 11 と、I/O システム 12 と、を備えて構成される。サブシステム 2 は、演算システム 21 と、I/O システム 22 と、を備えて構成される。

【0019】

演算システム 11, 21 には、同期した周波数 166 MHz のクロック信号 CLK が供給される。このように、演算システム 11, 21 に、同期したクロック信号 CLK が供給されることにより、サブシステム 1, 2 は、ロックステップ方式に従って、同一の処理を同期して同時に実行する。

【0020】

サブシステム 1, 2 の間には、分周器 31 が接続される。分周器 31 は、FSB のクロック信号 CLK が供給されて、このクロック信号 CLK を分周するものであり、クロック信号 CLK を分周して同期信号 S1 を生成する。

【0021】

分周器 31 は、生成した同期信号 S1 を演算システム 11 のメモリブリッジ 16、演算システム 21 のメモリブリッジ 26、I/O システム 12 の I/O ブリッジ 18、I/O システム 22 の I/O ブリッジ 28 に、それぞれ、供給する。

【0022】

尚、本実施の形態では、データの送受信に、PCI-Express インタフェースを用いるものとする。PCI-Express インタフェースでは、パラレルバスで生ずる信号線と信号線との間でのデータのずれ（スキュー）を防止するため、シリアルリンクが採用されている。演算システム 11, 21 と I/O システム 12, 22 との間も PCI-Express インタフェースに従って接続される。

【0023】

分周器 31 は、同期信号 S1 の 1 周期が、2.5Gbps/lane の PCI-Express インタフェースの 24 シンボルタイムに相当するように、クロック信号 CLK の周波数 166 MHz を 1/16 の周波数 10.4 MHz に分周する。

【0024】

PCI-Express インタフェースでは、演算システム 11, 21、I/O システム

12, 22といったデバイスは、1対1で接続される。差動信号を用いてデータを伝送する場合、リンクには、片方向に2本、双方向で合計4本の信号線が用いられる。この4本の信号線の組は、レーンと呼ばれるものである。

【0025】

また、1シンボルタイムとは、PCI-Expressに従って、1レーンのデータを8B/10Bエンコーディング処理した後の1バイトの有効データを送信するために必要な時間をいう。

【0026】

同期信号S1の周波数が10.4MHzになることにより、I/Oブリッジ18, 28、メモリブリッジ16, 26は、互いに、同期信号S1の1周期で、1レーンあたり24バイトの有効データを送信することができる。

【0027】

演算システム11は、プロセッサ13, 14と、主記憶装置15と、メモリブリッジ16と、を備えて構成される。また、演算システム21は、プロセッサ23, 24と、主記憶装置25と、メモリブリッジ26と、を備えて構成される。

【0028】

プロセッサ13, 14, 23, 24は、演算処理を実行するものである。主記憶装置15, 25は、データ等を記憶するものである。メモリブリッジ16とプロセッサ13, 14、メモリブリッジ26とプロセッサ23, 24とは、それぞれ、フロントサイドバス(FSB)を介して接続され、クロック信号CLKに同期して動作する。

【0029】

メモリブリッジ16, 26は、I/Oブリッジ18, 28とデータの送受信を行うものである。また、メモリブリッジ16, 26は、通信エラー信号S2を相互に送受信する。この通信エラー信号S2は、メモリブリッジ16, 26との間で、通信エラーの情報を共有して、連携してエラー処理を行うための信号であり、オープン・ドレイン信号として送受信される。尚、メモリブリッジ16, 26の詳細な構成については、後述する。

【0030】

I/Oシステム12は、I/Oデバイス17と、I/Oブリッジ18と、コンフィギュレーションレジスタ19と、を備えて構成される。I/Oシステム22は、I/Oデバイス27と、I/Oブリッジ28と、コンフィギュレーションレジスタ29と、を備えて構成される。

【0031】

I/Oデバイス17、27は、それぞれ、I/Oブリッジ18、28との間でデータの送受信を行うものである。

【0032】

I/Oブリッジ18、28は、それぞれ、I/Oデバイス17、27との間、又はメモリブリッジ16、26との間で、シリアル伝送を行うものである。

I/Oブリッジ18、28とメモリブリッジ16、26とは、PCI-Expressインタフェースのx8のリンクL1によって接続される。

【0033】

即ち、演算システム11、21のそれぞれのメモリブリッジ16、26と、I/Oシステム12、22のそれぞれのI/Oブリッジ18、28とは、リンクL1を介してクロスリンク接続されている。即ち、メモリブリッジ16は、I/Oシステム12、22に接続され、演算システム21のメモリブリッジ26は、I/Oシステム12、22に接続されている。

【0034】

このようにクロスリンク接続されることにより、演算システム11、21は、それぞれ、I/Oシステム12、22に対して通信可能であり、I/Oシステム12、22は、それぞれ、演算システム11、21に対しても通信可能になる。

【0035】

尚、PCI-Expressインタフェースでは、層単位のアップグレードを可能とするため、機能が階層化されている。そして、プロトコルが各階層毎に定義されている。

【0036】

PCI-Expressインタフェースでは、図2(a)に示すように、トランザクション層において、データにヘッダが付与されてトランザクション層パケットが生成

される。

【0037】

また、図2（b）に示すように、データリンク層において、トランザクション層パケットにシーケンス番号とCRC（Cyclic Redundancy Check）とのステータス情報が付加されてデータリンク層パケット（DLLP：Data Link Layer Packet）が生成される。

【0038】

そして、図2（c）に示すように、物理層において、データリンク層パケットにフレームデータが付加される。そして、このパケットが送受信される。

【0039】

I/Oブリッジ18、28は、このようなPCI-Expressインタフェースに従ってデータを送受信するためのインタフェース回路部（図示せず）を備える。

【0040】

コンフィギュレーションレジスタ19、29は、分周器31から供給された同期信号S1の1周期中において、I/Oブリッジ18、28が送信するアップストリームのパケットのパケット長、データ数を制限するためのデータを保持するものである。

【0041】

このようにパケット長、データ数に制限を設けたのは、送信されるパケットが、送信経路長、クロックのドリフトの影響を受けないようにするためである。具体的には、送信する各パケットの最大パケット長は、192バイトとされる。

【0042】

I/Oブリッジ18、28は、この制限に従い、同期信号S1の立ち上がりタイミングで、それぞれ、メモリブリッジ16、26に、同一のパケットを同時に送信する。

【0043】

I/Oブリッジ18、28は、複数の小さなパケットを送信する場合、同期信号S1の1周期において送信可能な最大データ数を越えないように、送信制御を行い、1つのパケットが、周波数10.4MHzの同期信号S1の1周期を越え

ないようにする。

【 0 0 4 4 】

尚、このコンフィギュレーションレジスタ 1 9 の値を B I O S (Basic Input/Output System) を用いて変更することができる。サブシステム 1 は、このような B I O S を記憶する不揮発性メモリ (図示せず) を備える。

【 0 0 4 5 】

このように構成されたコンピュータシステムは、システム間の通信内容の比較により障害診断を行う。また、このコンピュータシステムは、特定のシステムでの故障が判断されると該当するシステムをマスクして、残りのシステムによって実行中の処理を継続する。

【 0 0 4 6 】

次に、メモリブリッジ 1 6 , 2 6 の構成について説明する。尚、メモリブリッジ 2 6 はメモリブリッジ 1 6 と同様に構成されたものであり、ここでは、メモリブリッジ 1 6 の構成についてのみ説明する。

【 0 0 4 7 】

メモリブリッジ 1 6 は、図 3 に示すように、インタフェース回路部 4 0 と、同期化用バッファ 5 0 と、内部回路部 6 0 と、からなる。

【 0 0 4 8 】

インタフェース回路部 4 0 は、PCI-Express インタフェースに対応して設けられたものであり、データリンク/物理層 4 1 と、トランザクション層 4 2 と、に区分される。

【 0 0 4 9 】

また、データリンク/物理層 4 1 は、物理層 4 3 - 1 ~ 4 3 - n と、データリンク層 (R X) 4 4 と、データリンク層 (T X) 4 5 と、に区分され、トランザクション層 4 2 は、通信エラー処理部 4 6 と、トランザクション層 4 7 と、に区分される。

【 0 0 5 0 】

このデータリンク/物理層 4 1 、トランザクション層 4 2 、内部回路部 6 0 は、互いに異なるクロック信号に同期して動作する。

【0051】

物理層 43-1～43-n は、同期信号 S2 の 1 周期中において、図 2 (c) に示すようなパケットを送受信するための層である。インタフェース回路部 40 は、送受信するパケットを保持するためのエラスティックバッファ (EB) を備える。尚、インタフェース回路部 40 は、物理層 43-1～43-n において、通信エラーを検出するとエラー情報を出力する。

【0052】

データリンク層 (RX) 44 は、図 2 (c) に示すパケットからデータリンク層パケットを取り出すための層である。

【0053】

データリンク層 (TX) 45 は、通信エラー処理部 46 が出力した ACK/NACK/f low 制御信号を受け取るための層である。

【0054】

通信エラー処理部 46 は、通信エラーの処理を行うものである。

従来の PCI-Express では、データリンク層 (RX) 44 が、ステータス情報のうち、いくつかのエラー信号を、トランザクション層 47 に、また、ACK/NACK/f low 制御信号をデータリンク層 (TX) 45 に直接供給する。

【0055】

しかし、本実施の形態では、通信エラー処理部 46 が備えられ、通信エラー処理部 46 がこれらのステータス情報を、データリンク層 (RX) 44 において取得する。そして、通信エラー処理部 46 が取得したステータス情報をトランザクション層 47、データリンク層 (TX) 45 に供給するように構成されている。

【0056】

通信エラー処理部 46 は、データリンク層パケットに付加された CRC をチェックして、通信エラーを検出し、エラー情報を出力する。

【0057】

通信エラー処理部 46 は、物理層 43-1～43-n またはデータリンク層 (RX) 44 において、通信エラーが検出されなければ、トランザクション層 47 とデータリンク層 (TX) 45 とに、データ、ステータス情報をそのまま出力す

る。インタフェース回路部 40 は、受信したデータに通信エラーがなければ、ステータス情報に従って、データを送信した I/Oブリッジ 18, 28 に定期的に ACK 信号を返送する。

【0058】

一方、通信エラー処理部 46 は、物理層 43-1 ~ 43-n またはデータリンク層 (RX) 44 において、通信エラーが検出されると、同期信号 S1 の 1 周期の間に受信した全パケットをロストパケットとしてキャンセルし、受信データのトランザクション層 47 への出力を停止する。

【0059】

通信エラー処理部 46 は、パケットをキャンセルすると、データリンク層 (RX) 44 に対して、次に受信する予定のパケットのシーケンス番号を、通信エラーパケット受信前のシーケンス番号に設定するように指示する。

【0060】

また、通信エラー処理部 46 は、通信エラーを検出すると、同期信号 S1 の 1 周期の間、通信エラー信号 S2 をアサート (assert)、即ち、有効とする。通信エラー処理部 46 は、アサートした通信エラー信号 S2 を、信号線を介してメモリブリッジ 26 に送信する。

【0061】

トランザクション層 47 は、上位のソフトウェア層からの読み出しと書き込み要求を受け付けるとともにデータリンク層 (RX) 44、データリンク層 (TX) 45 に対してパケットの転送を要求するための層である。

【0062】

同期化用バッファ 50 は、トランザクション層 47 と内部回路部 60 との間で、データの受け渡しをするためのバッファであり、トランザクション層 47 から出力されたデータを保持する。

【0063】

内部回路部 60 は、同期化用バッファ 50 が保持しているデータを同期信号 S1 に同期したタイミングで取得し、取得したデータをプロセッサ 13, 14、主記憶装置 15 に送出するための回路である。

【0064】

尚、I/Oブリッジ18、28がメモリブリッジ16、26にシリアルデータを送信する場合、I/Oブリッジ18、28が送信インタフェースとなり、メモリブリッジ16、26のインタフェース回路部が受信インタフェース部となる。

【0065】

また、前述のようにメモリブリッジ16、26がI/Oブリッジ18、28にシリアルデータを送信することも可能であり、この場合、メモリブリッジ16、26が送信インタフェースとなり、I/Oブリッジ18、28のインタフェース回路部が受信インタフェース部となる。

【0066】

次に本実施の形態に係るコンピュータシステムの動作を説明する。

尚、ここでは、I/Oブリッジ18が、メモリブリッジ16、26にシリアルデータを送信する場合について説明する。

【0067】

I/Oブリッジ18は、I/Oデバイス17からデータが供給されると、I/Oブリッジ18は、図2（a）に示すように、トランザクション層において、シリアルデータにヘッダを付与し、トランザクション層パケットを生成する。

【0068】

I/Oブリッジ18は、図2（b）に示すように、データリンク層において、生成したトランザクション層パケットに、シーケンス番号とCRCとのステータス情報を付加し、データリンク層パケットを生成する。

【0069】

そして、I/Oブリッジ18は、図2（c）に示すように、物理層において、生成したデータリンク層パケットに、フレームデータを付加する。そして、I/Oブリッジ18は、図2（c）に示すパケットを、リンクL1を介してメモリブリッジ16、26に送信する。

【0070】

メモリブリッジ16のインタフェース回路部40は、物理層43-1～43-nにおいて、このデータを受信する。

インタフェース回路部 40 は、図 4 (a) に示すように、物理層 43-1 ~ 43-n において、同期信号 S1 の 1 周期中において受信した全パケットを、エラストックバッファに、一旦、格納してから、データリンク層 (RX) 44 に出力する。

【0071】

インタフェース回路部 40 は、データリンク層 (RX) 44 において、図 2 (c) に示すパケットからデータリンク層パケットを取り出す。また、インタフェース回路部 40 は、図 2 (b) に示すデータリンク層パケットに含まれている CRC に基づいてエラーの検出を行う。

【0072】

通信エラー処理部 46 は、図 4 (c) に示すように、同期信号 S1 の各周期中において受信したパケットを、それぞれ、次の同期信号 S1 の立ち上がり同期して、取得する。

【0073】

受信したパケットにエラーが検出されなければ、通信エラー処理部 46 は、図 4 (d) に示すように、通信エラー信号 S2 をハイ (H) レベルにしてディASSERT、即ち、無効とする。従って、受信されたパケットは有効となる。

【0074】

そして、通信エラー処理部 46 は、図 4 (e) に示すように、各パケットを、次の同期信号 S1 の立ち上がり同期してトランザクション層に送出する。

【0075】

インタフェース回路部 40 は、トランザクション層において、データリンク層パケットから、トランザクション層パケットを取り出し、さらに、データと取り出して、図 4 (f) に示すように同期化用バッファ 50 に出力する。

【0076】

内部回路部 60 は、同期信号 S1 の立ち上がり同期して、同期化用バッファ 50 からデータを取得し、取得したデータをプロセッサ 13, 14、主記憶装置 15 に出力する。

【0077】

I/Oブリッジ18からメモリブリッジ16, 26にデータを送信する場合、I/Oブリッジ18とメモリブリッジ16との間と、I/Oブリッジ18とメモリブリッジ26との間で、リンクL1の配線長に差異がほぼなければ、図5(b), (c)に示すように、メモリブリッジ16, 26は、ほぼ同時にデータを受信する。

【0078】

しかし、I/Oブリッジ28とメモリブリッジ26との間のリンクL1が、I/Oブリッジ18とメモリブリッジ16との間のリンクL1よりも長いと、図5(d), (e)に示すように、メモリブリッジ16, 26がデータを受信するタイミングに差異が生じる。

【0079】

但し、差異が生じても、同じ同期信号S1の周期内であれば、演算システム11, 21は、クロック信号CLKに同期して同じ処理を実行する。

もし、メモリブリッジ26が、同期信号S1の1周期目と2周期目とでデータを受信するようになるのであれば、I/Oブリッジ18は、1パケットにおけるデータの長さを短くするように、BIOSを用いてコンフィギュレーションレジスタ19が格納しているデータを変更する。

【0080】

次に、図6(a), (b)に示すように、メモリブリッジ16のインタフェース回路部40が、同期信号S1の第2クロックサイクルにおいて、受信したパケットの通信エラーを検出するものとする。

【0081】

この場合、通信エラー処理部46は、図6(c), (e)に示すように、第3クロックサイクルにおいて、データリンク層パケット(DLLP)をその一部に含んでいたとしても、すべてのパケットをキャンセル扱いとする。

【0082】

そして、通信エラー処理部46は、第3クロックサイクル以降のパケットもキャンセルする。通信エラー処理部46は、第2クロックサイクルでキャンセルしたパケットが再送されるまで、すべてのパケットの受信をキャンセルする。

【 0 0 8 3 】

通信エラー処理部 4 6 は、通信エラーが検出されると、データリンク層（R X） 4 4 において、管理されているパケットのシーケンス番号をエラー発生前の番号に設定し直す。

【 0 0 8 4 】

また、通信エラーが検出されると、通信エラー処理部 4 6 は、図 6（d）に示すように、通信エラー信号 S 2 をロー（L）レベルとしてアサート、即ち、有効とする。同期信号 S 1 の第 4 クロックサイクルになると、パケットは受信されないため、通信エラー処理部 4 6 を、ディアサートする。

【 0 0 8 5 】

通信エラー処理部 4 6 は、データの送信元である I/Oブリッジ 1 8 に、データの再送を要求する。I/Oブリッジ 1 8 は、メモリブリッジ 1 6 から再送要求を受信した場合、再送要求のあったパケットを再送する。また、メモリブリッジ 1 6 からの ACK 信号が返送されずに、所定期間が経過した場合にも、送信が確認されていないパケットの再送信を行う。

【 0 0 8 6 】

続いて、図 7（b）に示すように、メモリブリッジ 1 6 が、同期信号 S 1 の第 6 クロックサイクルにおいて、再送要求に応答してシーケンス番号 2 のパケットを受信し、通信エラーがなければ、インタフェース回路部 4 0 は、図 6（c）～（e）に示すように、キャンセルされた第 3 クロックサイクル以降のパケットを、そのまま、受信する。

【 0 0 8 7 】

次に、図 8（a）～（c）に示すように、メモリブリッジ 1 6 が受信したデータに通信エラーは検出されなくても、メモリブリッジ 2 6 が通信エラーを検出すると、メモリブリッジ 2 6 は、図 8（d）に示すように、メモリブリッジ 1 6 にローレベルの通信エラー信号 S 2 を出力する。

【 0 0 8 8 】

メモリブリッジ 2 6 は、同期信号 S 1 の第 3 クロックサイクルにおいて、通信エラー信号 S 2 がアサートされると、通信エラー処理部 4 6 は、図 8（e）に示

すように、同期信号 S 1 の第 3 クロックサイクルにおいて保持しているシーケンス番号 2 のパケットをキャンセルする。

【0089】

通信エラー処理部 46 は、同期信号 S 1 の第 4 クロックサイクル以降、トランザクション層 47 へのパケットの引き渡しを停止する。

そして、通信エラー処理部 46 は、データリンク層 (R X) 44 において、管理する次回受信予定パケットのシーケンス番号を、パケットキャンセル前の値に再設定させる。

【0090】

図 9 (a), (b) に示すように、メモリブリッジ 16 が、通信エラーを有するパケットに続いて、データリンク層パケットのみで構成されるパケットを受信した場合、通信エラー処理部 46 は、図 6 (c), (e) に示すように、まず、通信エラーを有するパケットをキャンセルする。

【0091】

通信エラー処理部 46 は、通信エラーを有するパケットをキャンセルすると、図 9 (d) に示すように、通信エラー信号 S 2 をアサートして、シーケンス 2 以降のパケット列の再送信を要求する。しかし、通信エラー処理部 46 は、通信エラーを有するパケットをキャンセルしても、図 6 (c), (e) に示すように、同期信号 S 1 の第 3 クロックサイクルにおいて受信されたデータリンク層パケットをキャンセルしない。これは、データリンク層パケットには、シーケンス番号がないため、シーケンス番号エラーが発生しないためである。

【0092】

エラー処理部 46 は、データリンク層パケットをキャンセルしなくても、再送されたパケットのシーケンス番号から、順序を特定することができ、メモリブリッジ 16 は、問題なく再送信されたパケットを受信することができる。

【0093】

以上説明したように、本実施の形態によれば、メモリブリッジ 16 のインタフェース回路部 40 が通信エラーを検出した場合、通信エラー処理部 46 が受信したパケットをキャンセルする。そして、通信エラー処理部 46 は、アサートした

通信エラー信号 S 2 をメモリブリッジ 2 6 に出力し、キャンセルしたパケットの再送をパケット送信元に要求するようにした。

【0094】

従って、通信エラーが生じた場合でも、メモリブリッジ 1 6, 2 6 の通信エラー処理部が連携してパケット送信元に、パケットの再送信をするようになるため、受信データの同期ずれを回避することができる。そして、演算システム 1 1, 2 1 は、同一のデータ処理を同期して行うことができる。

【0095】

また、I/Oブリッジ 1 8 が、リンク L 1 の配線長に差異が生じても、影響を受けないように、送信するパケットのパケット長、データ数に制限が設けられるようにした。

【0096】

従って、フォールトトレラントコンピュータシステムを構成するための回路基盤設計および筐体設計が容易になる。

【0097】

尚、本発明を実施するにあたっては、種々の形態が考えられ、上記実施の形態に限られるものではない。

例えば、上記実施の形態では、メモリブリッジ 1 6, 2 6、I/Oブリッジ 1 8, 2 8 に、インタフェース回路部を備えるようにした。しかし、図 1 0 に示すように、演算システム 1 1, 2 1 において、メモリブリッジ 1 6, 2 6 とは別に、それぞれ、送受信ブリッジ 7 1, 7 2 を備え、また、I/Oシステム 1 2, 2 2 とは別に、それぞれ、送受信ブリッジ 8 1, 8 2 を備えることができる。

【0098】

この場合、送受信ブリッジ 7 1, 7 2, 8 1, 8 2 は、それぞれ、通信エラー処理部を備える。送受信ブリッジ 7 1, 7 2 は、それぞれ、メモリブリッジ 1 6, 2 6 に接続され、送受信ブリッジ 8 1, 8 2 は、それぞれ、I/Oブリッジ 1 8, 2 8 に接続される。

【0099】

そして、送受信ブリッジ 7 1, 7 2 と、送受信ブリッジ 8 1, 8 2 とは、互い

にロックステップ方式に従って同期化されてデータの送受信を行う。尚、送受信ブリッジ 71, 72 は、既存のメモリブリッジ 16, 26 とは、1 セットの通信リンクで接続される。接続は、既存のメモリブリッジのサポートする高速シリアルリンクによって行われる。この場合、メモリブリッジ 16 と送受信ブリッジ 71、メモリブリッジ 26 と送受信ブリッジ 72 との間のリンクの配線長は、受信タイミングの差異による通信エラーの発生を避けるため、できるだけ、短くされる。

【0100】

このように構成されることにより、既存のシステムチップセットコンポーネントをそのまま利用して、フォールトトレラントコンピュータシステムを構成することができる。

【0101】

本実施の形態では、コンピュータシステムに 2 つのサブシステム 1, 2 を備え、また、サブシステム 1, 2 に、それぞれ、2 つのプロセッサ 13, 14, 23, 24 を備えて、コンピュータシステムを二重化冗長構成とした。しかし、コンピュータシステムの構成は、これに限られるものではなく、三重化冗長構成、あるいはそれ以上の冗長構成とすることができる。

【0102】

また、本実施の形態では、高速シリアルリンクとして PCI-Express を例として説明した。しかし、リンクとしては、これに限られるものではなく、InfiniBand、HyperTransport 等の他の高速シリアルリンクを用いることもできる。

【0103】

また、本実施の形態では、メモリブリッジ 16, 26 と、I/Oブリッジ 18, 28 との間で送受信するデータをシリアルデータとして説明したが、パラレルデータであっても、本実施の形態を適用することができる。

【0104】

【発明の効果】

以上説明したように、本発明によれば、通信エラーが生じた場合でも同一のデータ処理を同期して行うことができる。

【図面の簡単な説明】**【図 1】**

本発明の実施の形態に係るコンピュータシステムの構成を示すブロック図である。

【図 2】

送受信されるパケットデータの構成を示す説明図である。

【図 3】

図 1 に示すメモリブリッジの詳細な構成を示すブロック図である。

【図 4】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 5】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 6】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 7】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 8】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 9】

図 1 に示すメモリブリッジの動作を示すタイミングチャートである。

【図 1 0】

本発明の実施の形態に係るコンピュータシステムの応用例を示すブロック図である。

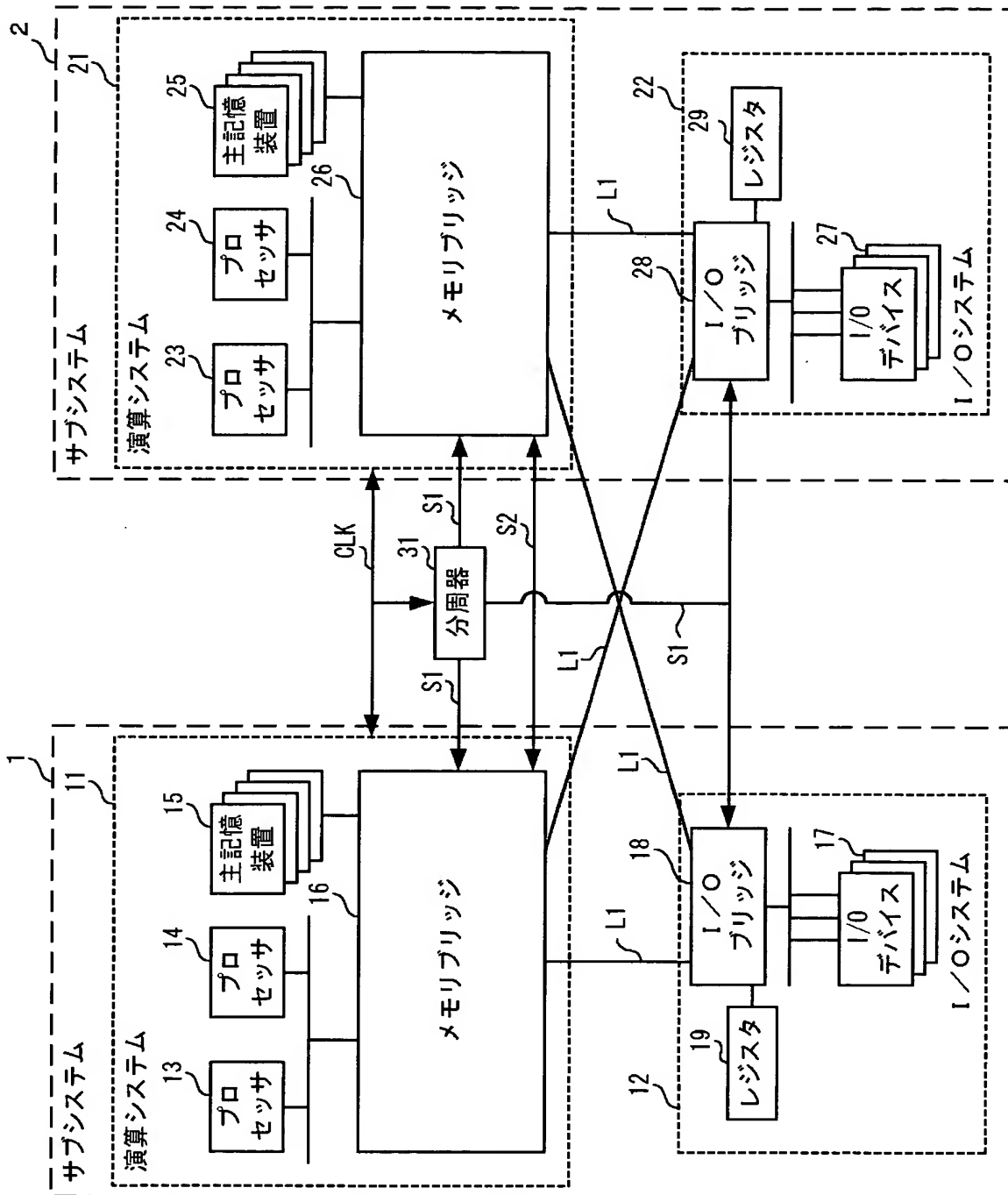
【符号の説明】

- 1, 2 サブシステム
- 1 3、1 4, 2 3, 2 4 プロセッサ
- 1 6, 2 6 メモリブリッジ
- 1 8, 2 8 I/Oブリッジ
- 4 6 通信エラー処理部

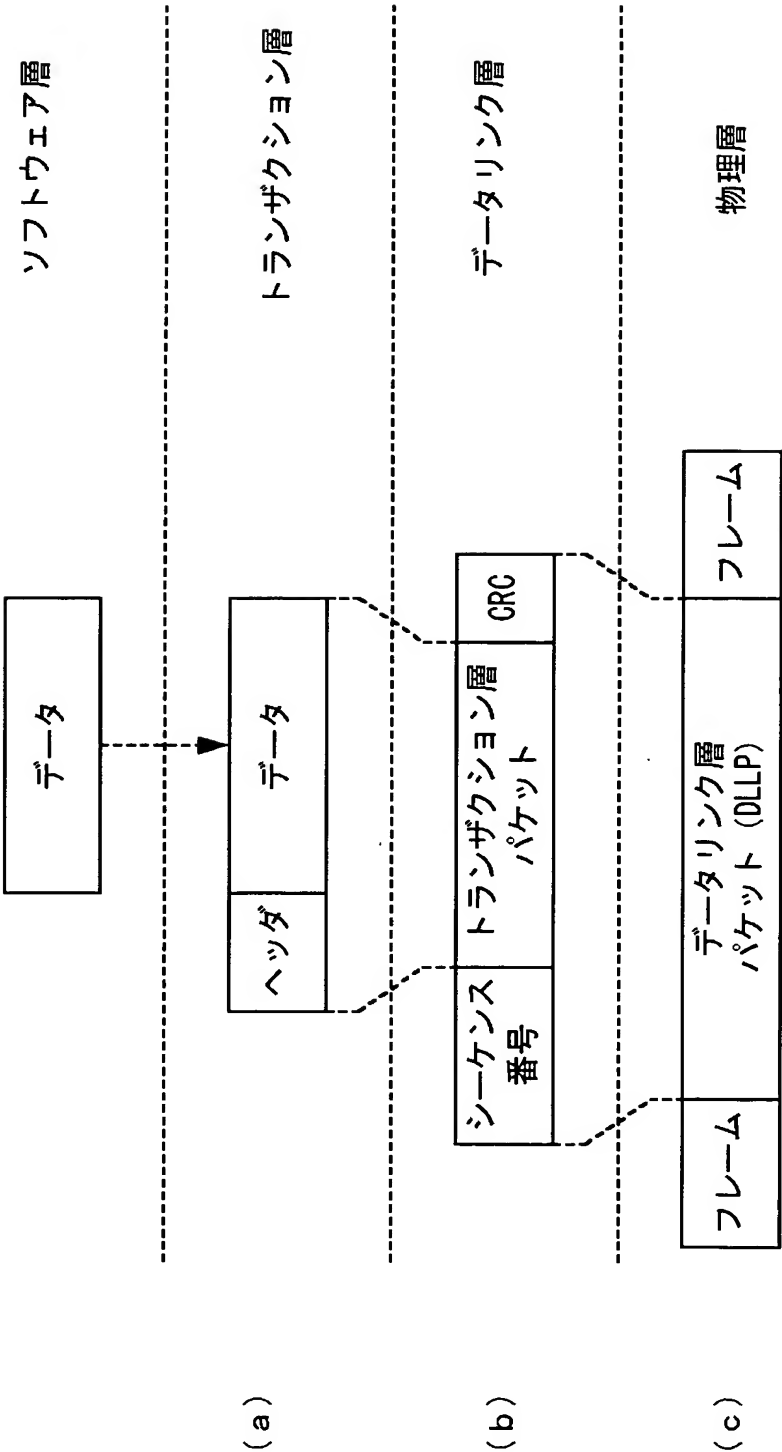
3 1 分周器

【書類名】 図面

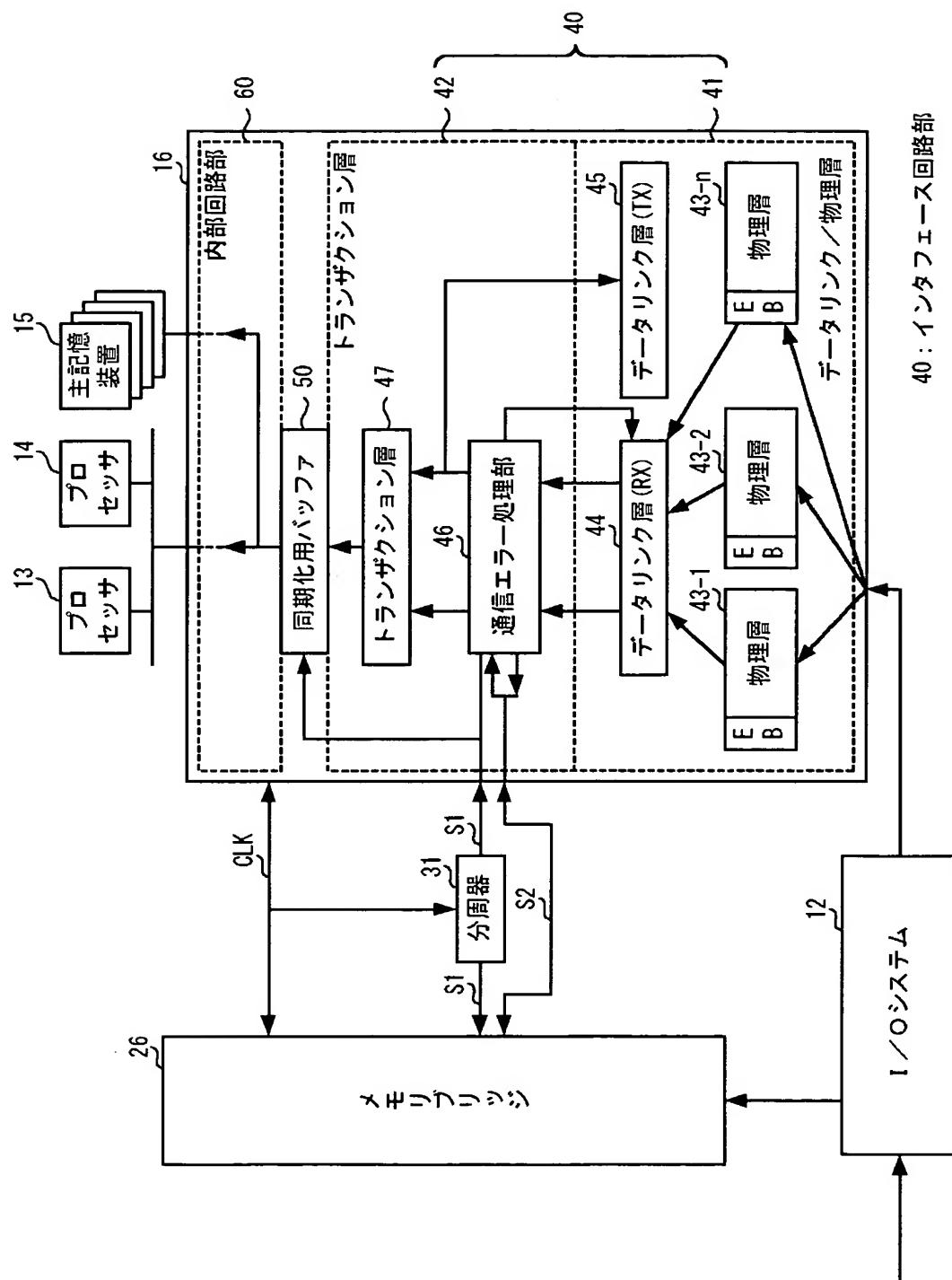
【図 1】



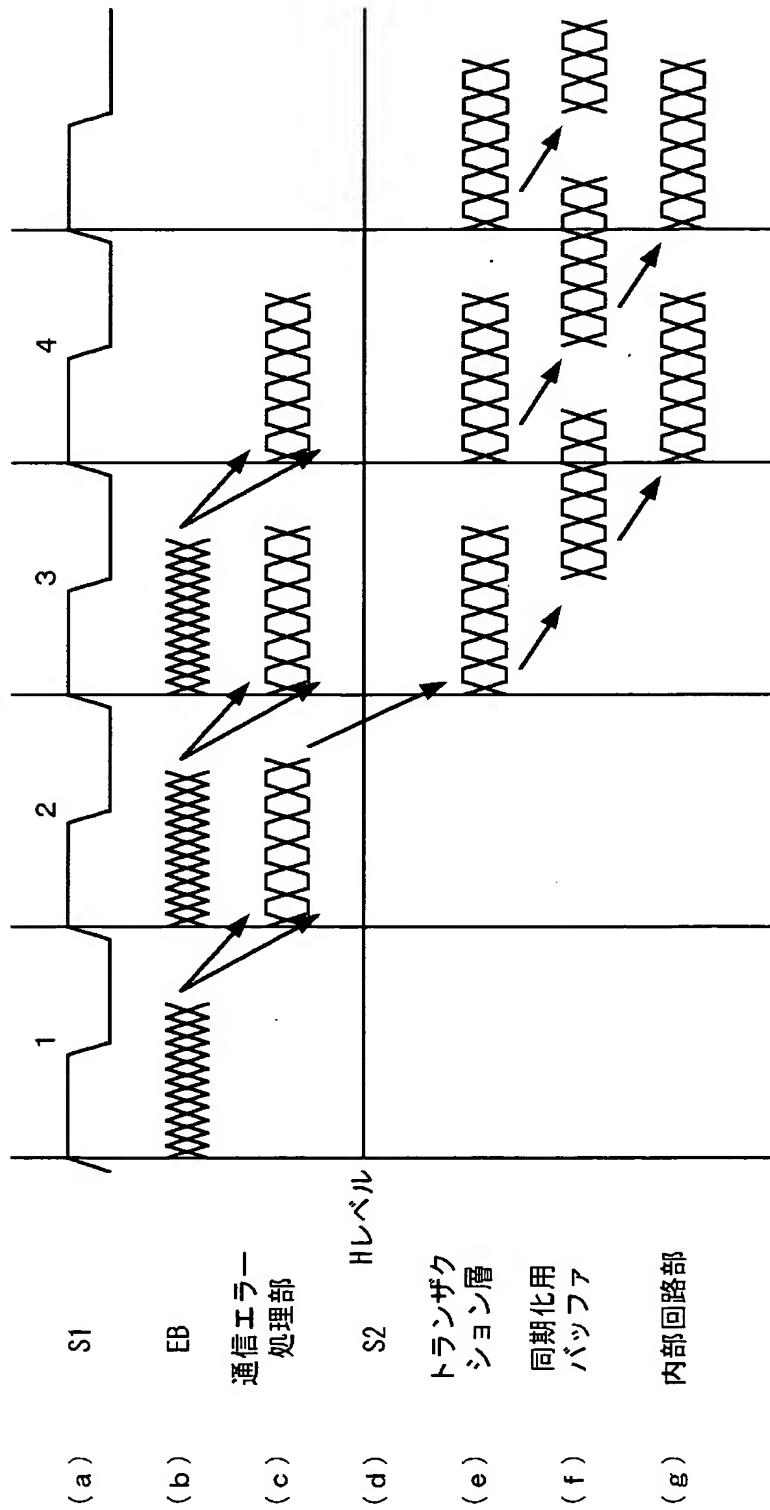
【図 2】



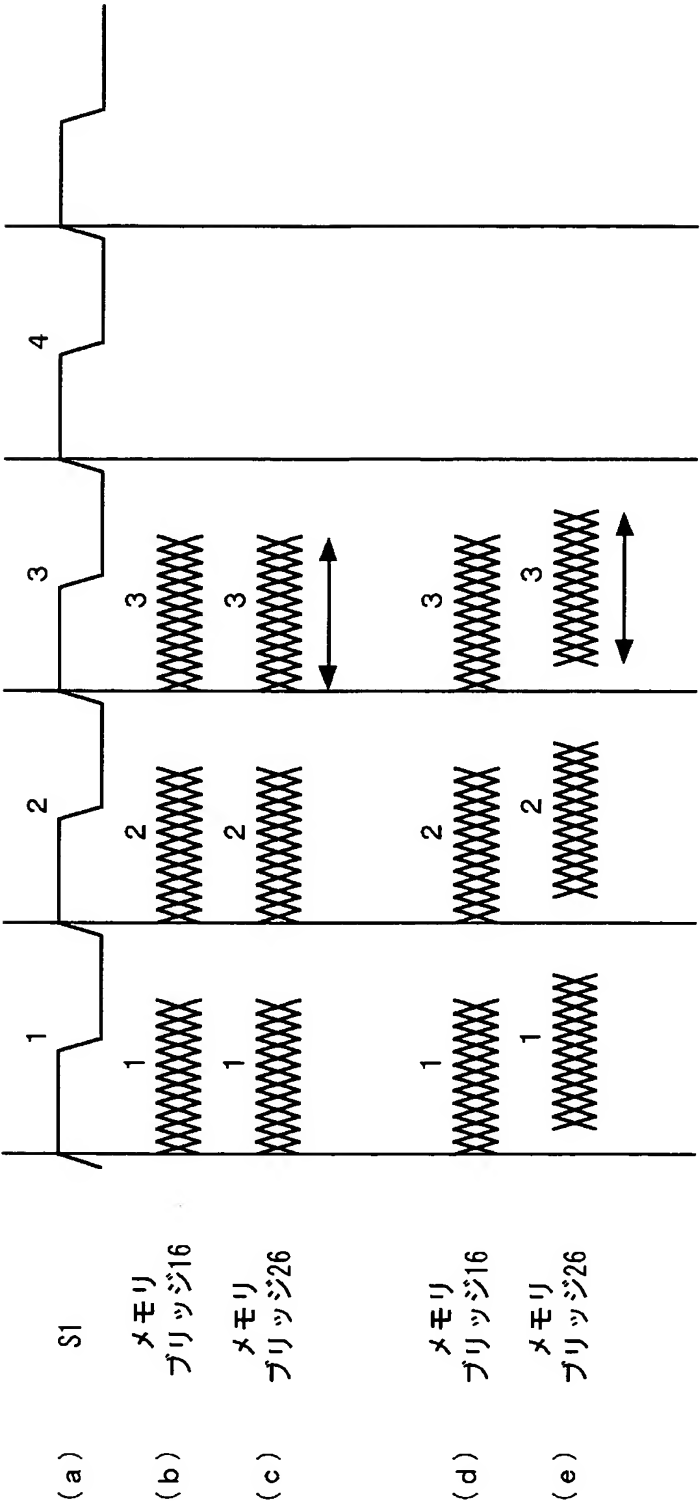
【図 3】



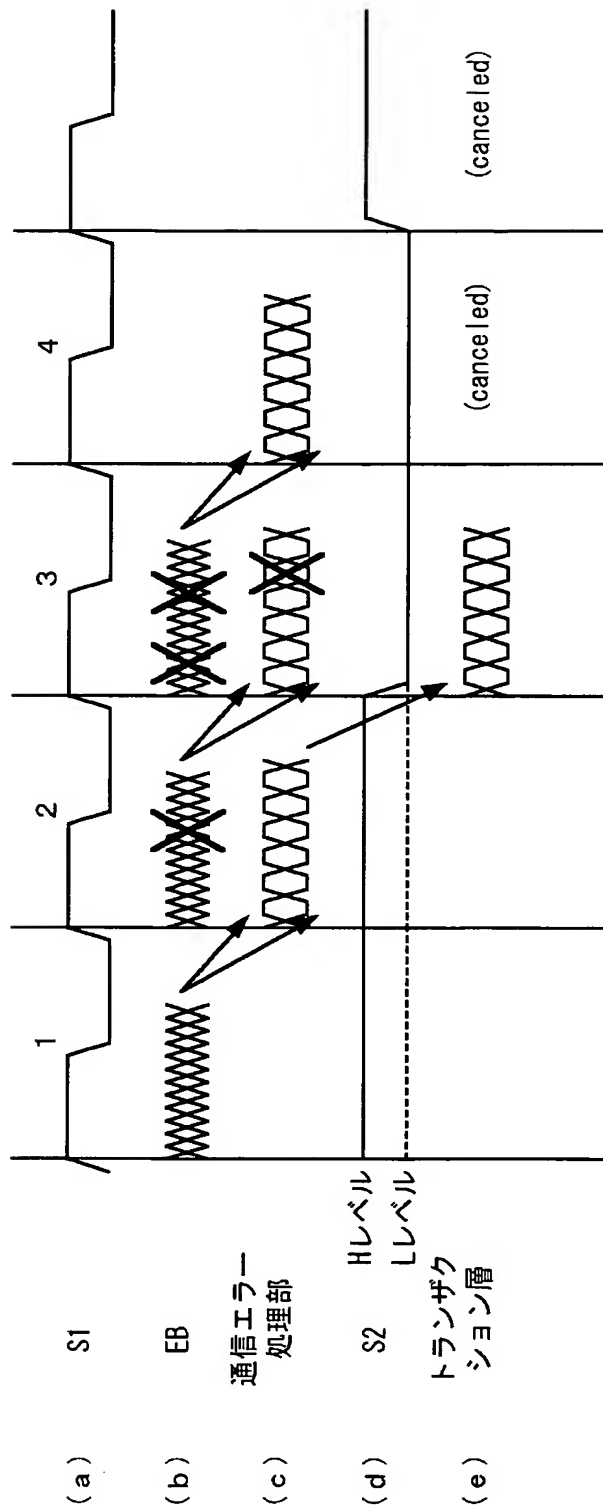
【図 4】



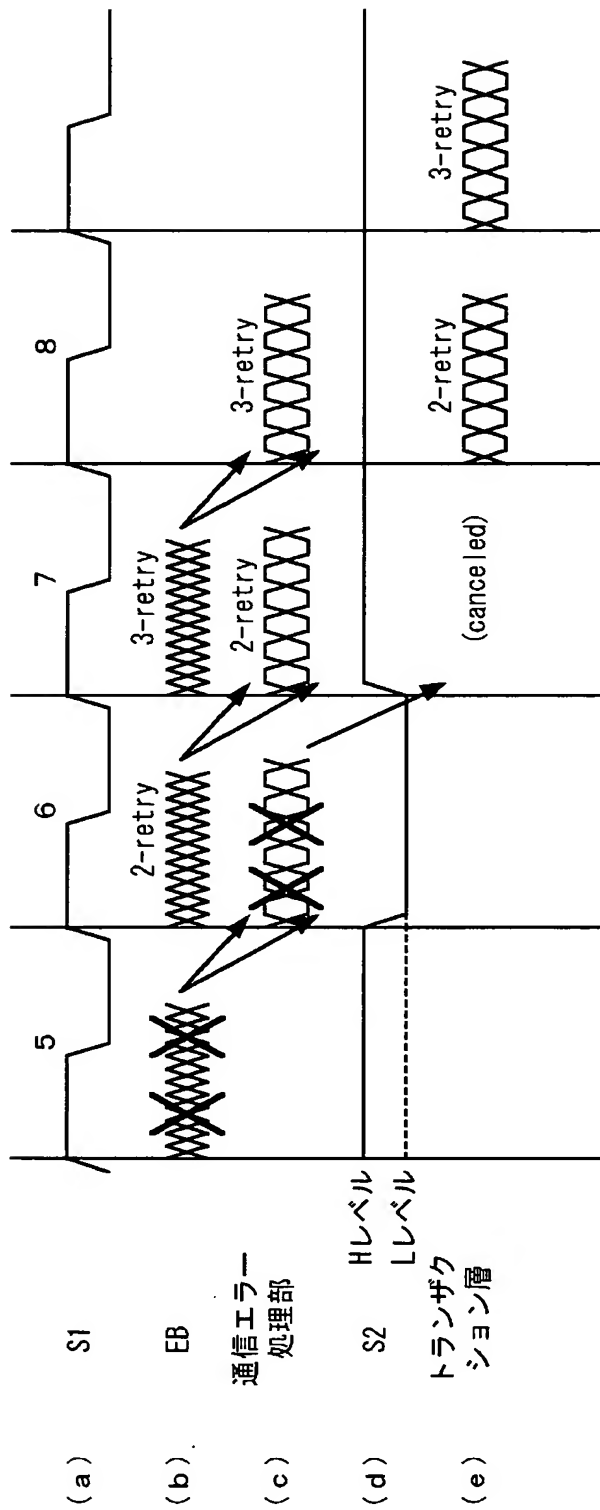
【図 5】



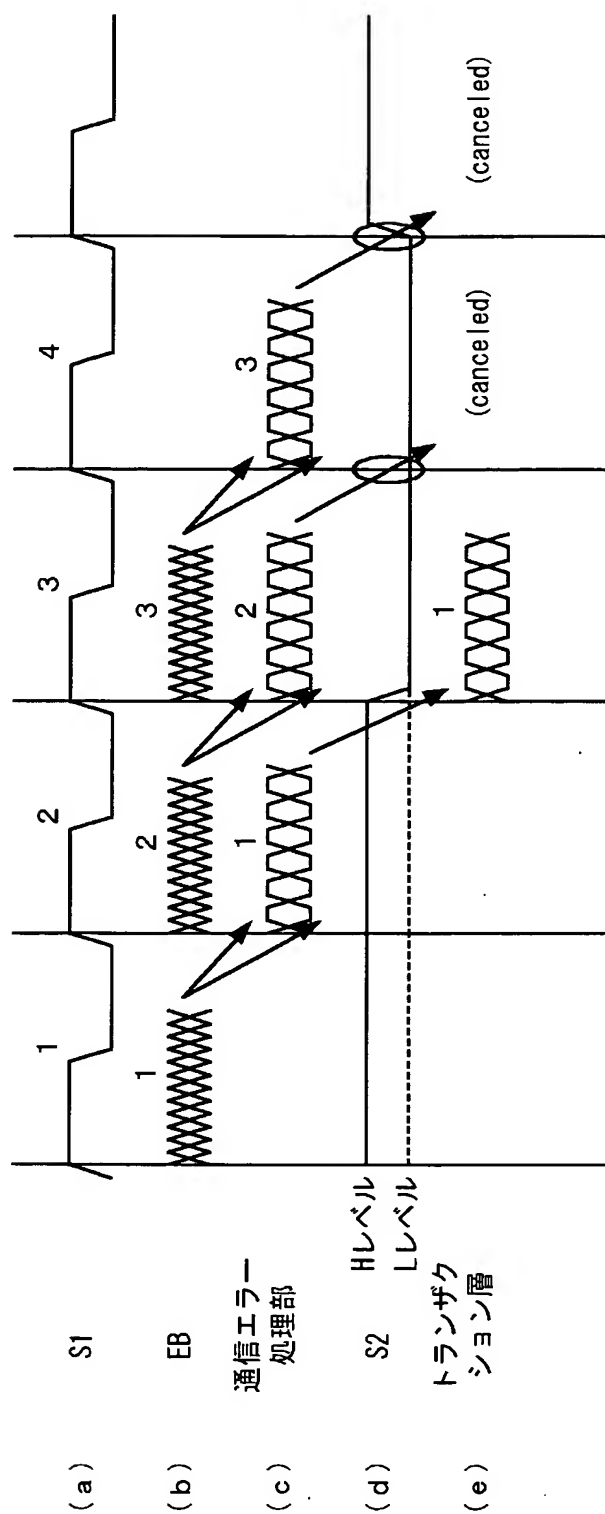
【図 6】



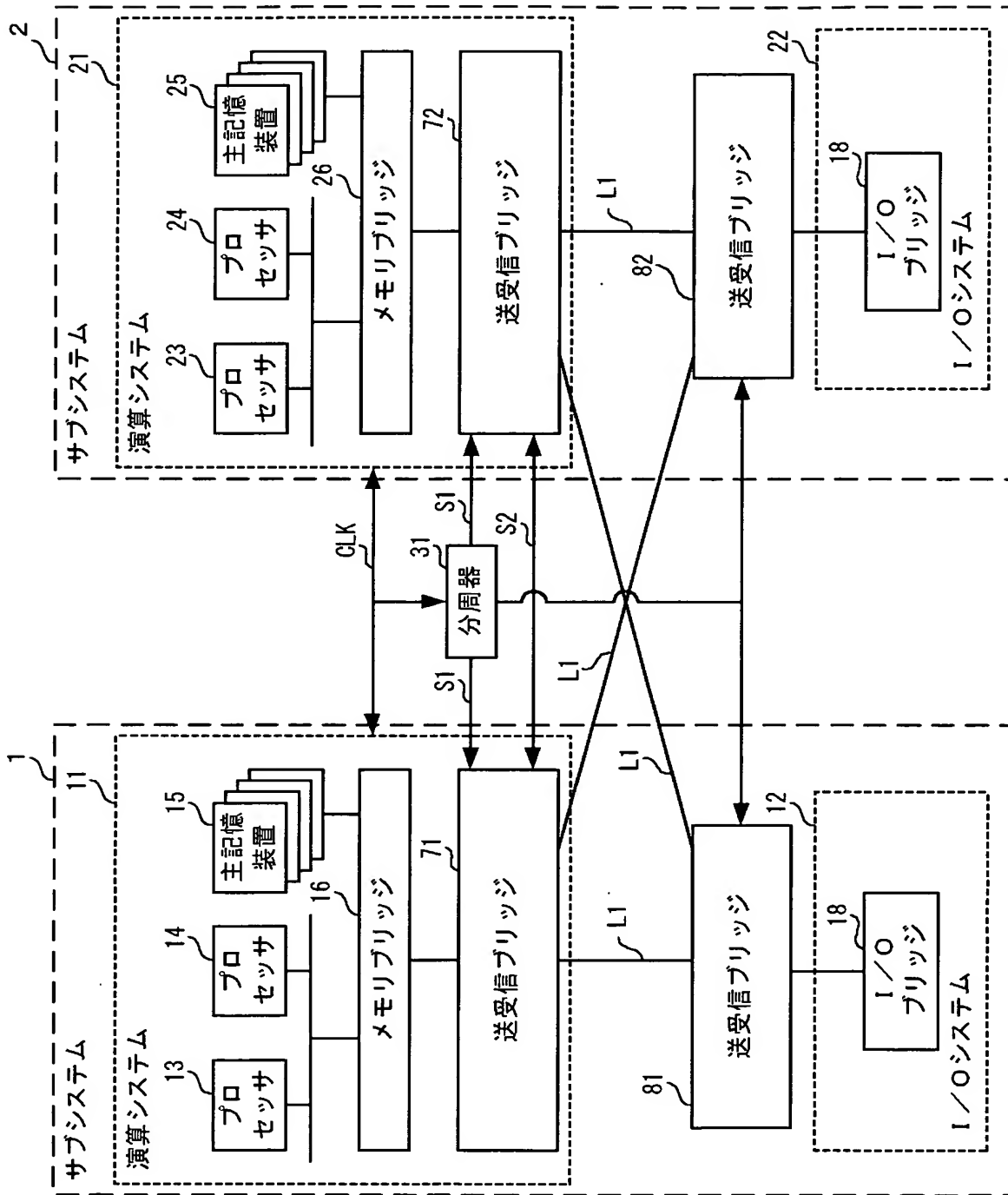
【図 7】



【図 8】



【図 10】



【書類名】 要約書

【要約】

【課題】 通信エラーが生じた場合でも同一のデータ処理を同期して行えるようにする。

【解決手段】 クロスリンク接続されたメモリブリッジ 1 6, 2 6 と、I/Oブリッジ 1 8, 2 8 とには、PCI-Express インタフェースに従ってデータの送受信を行うためのインタフェース回路部が、それぞれ、備えられる。また、各インタフェース回路部には、通信エラー処理部が備えられる。メモリブリッジ 1 6 の通信エラー処理部は、I/Oブリッジ 1 8 から受信したデータにエラーが発生すると、受信したデータをキャンセルしてメモリブリッジ 2 6 に通信エラー信号を出力する。メモリブリッジ 2 6 は、この通信エラー信号が供給されてデータの受信を停止する。そして、メモリブリッジ 1 6 の通信エラー処理部は、I/Oブリッジ 1 8 にデータの再送信を要求する。

【選択図】 図 1

特願 2 0 0 3 - 1 1 5 6 2 1

出 願 人 履 歴 情 報

識別番号 [0 0 0 0 0 4 2 3 7]

1. 変更年月日	1 9 9 0 年 8 月 2 9 日
[変更理由]	新規登録
住 所	東京都港区芝五丁目 7 番 1 号
氏 名	日本電気株式会社